Efficient, high sensitivity detection of oncogenic variants with UMIs and target enrichment

Michael Bocek, Lydia Bonar, Jean Challacombe, Richard Gantt, Rebecca Liao, Derek Murphy and Esteban Toro

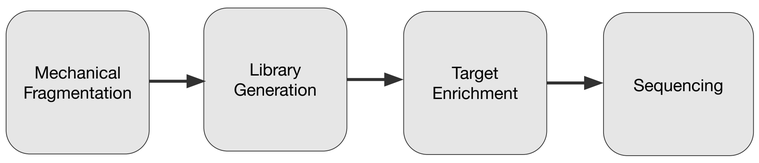
TWIST BIOSCIENCE

1. Abstract

Background/objectives:
Early detection can significantly improve clinical outcomes for a number of cancers, but many of the best current screening methods require invasive procedures. A promising alternative approach is a liquid biopsy of cell-free DNA (cfDNA) from plasma. Because tumors generally shed relatively large amounts of DNA into the circulation, cancer can potentially be detected by identifying oncogenic variants in cfDNA. This process generally requires extremely deep sequencing, and in many cases is limited by the accuracy of next-generation sequencing (NGS).
Methods:
One approach to overcoming this limitation is the use of unique molecular identifiers (UMIs), short sequences that uniquely tag each input DNA molecule prior to preparing NGS libraries. The approach can further be improved by tagging each original strand of the DNA molecule, in a technique termed duplex sequencing, which can correct early PCR errors and/or single-strand DNA damage events. Here we describe a new library preparation system incorporating short, discrete UMI sequences to maximize sequence distances for error correction.
Results:
We show that this system can determine the conversion efficiency of NGS libraries. Using the Twist cfDNA Pan-cancer Reference Standards to simulate a low fraction of tumor DNA in a healthy background, we demonstrate high sensitivity towards a variety of oncogenic substitutions, indels and structural variants. We demonstrate the baseline error rate using unmodified human cfDNA, and use the system to determine the mutation frequency in a synthetic biology application.
Conclusion:
In summary, this study demonstrates the utility of UMIs for a variety of applications in NGS.

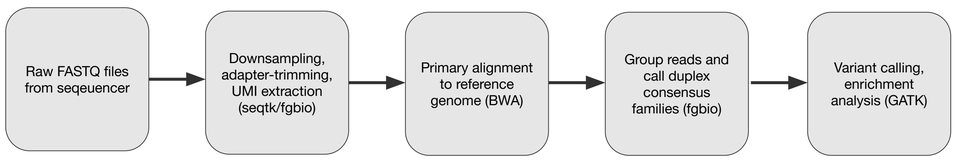
2. Methods

Experimental methods:



Human gDNA (NA12878, Coriell) or DNA constructs were fragmented on a Covaris ML230. NGS libraries were generated with fragmented gDNA or cfDNA (Twist cfDNA Pan-cancer Reference Standards) using the Twist Mechanical Fragmentation Library Preparation kit and Twist UMI Adapter System. Unless otherwise specified, hybrid capture was performed with target enrichment panels of various sizes (and 500 ng of DNA per library following the Twist recommendations for 16-hour hybridization reactions. Sequencing was performed with a NextSeq® 500/550 High Output v2 kit to generate 2x76 paired end reads.

Bioinformatics methods:



After base calling and FASTQ generation, reads were first downsampled to a fixed depth based on the target space of the panel. Reads were then pre-processed to mark adapter sequences (Picard) and to isolate UMI sequences (fgbio) into an unaligned BAM file. Raw reads were aligned to the human reference genome (hg38/GRCh38) using BWA, and were merged with the unaligned BAM to provide UMI information. After alignment, UMIs were error-corrected and grouped based on strand and UMI sequence, and consensus reads were called with a duplex strategy (fgbio). Unless otherwise specified, reads were subsequently filtered to keep only duplex consensus families, or those with at least one supporting read derived from each strand. After consensus calling, raw allele counts were obtained using samtools, or variant calls were obtained using Mutect2 (GATK) depending on the specific needs.

3. Design of UMI-containing adapters

To facilitate efficient sequencing and error correction, UMI sequences were derived from a discrete pool of 32 sequences. Sequences were chosen to have acceptable GC content, avoid homopolymers, and to have sufficient sequence distance between them (Hamming distance of at least 2) to allow for error detection even with a large number of UMI families present. Another key goal in sequence selection was the tradeoff between sequence length and complexity - error correction is easier in longer sequences, but requires a larger number of read cycles to UMI sequencing, and presents additional challenges for target enrichment.

To illustrate how a system of discrete UMI sequences can aid in error correction, the full range of possible sequencing errors within the UMI was simulated for different numbers of coincidental families. With a hamming distance of 2, all errors are detectable but not necessarily correctable in the UMI sequence. However with the number of distinct molecules per family, unresolvable single sequence errors should be rare. This was confirmed (figure 1A), showing that even with 30 coincidental coordinate pairs, over 90% of single-base substitution errors are correctable in most cases, as they do not produce sequences with a hamming distance of 1 to multiple other UMI sequences in the pool. Given the low-rate of mutations within the UMI, sequencing errors are unlikely to cause collisions between UMI sequences.

To confirm that the system was robust even to high input quantities of genomic DNA, a titration experiment comparing the number of large (>20 distinct molecules sharing a single set of mapping coordinates) coordinate pairs across a range of mass inputs and mapping densities was performed. The majority of families (>99%) have fewer than 20 distinct members, confirming that 1024 pairs of UMIs effectively samples the available diversity (Figure 1B).

To confirm that UMI sequences are fully compatible with target enrichment, we compared coverage over several different capture panels using either standard Twist Universal or UMI-containing adapters. No major differences were seen in performance between the two adapter sets (Figure 1C).

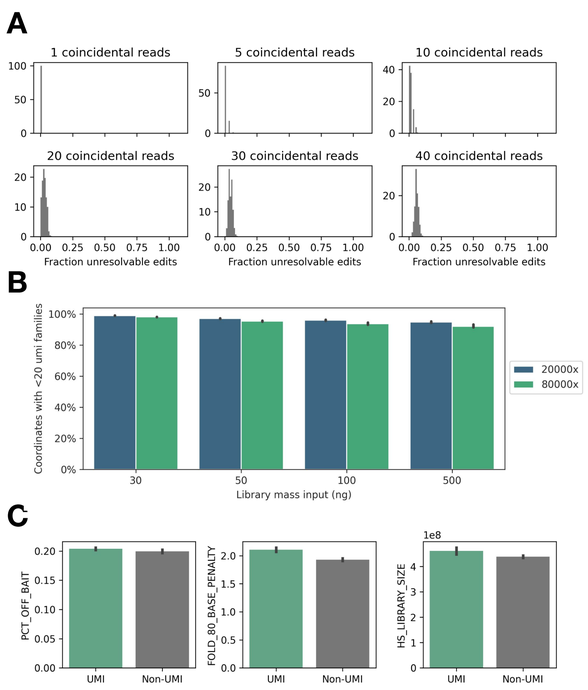


Figure 1: **(A)** Fraction of single-base edits that may give rise to ambiguous UMI sequences for family sizes of 1, 5, 10, 20, 30 and 40 reads. **(B)** Fraction of families under 20 members for different library masses at 20,000x and 80,000x coverage. **(C)** Comparison of capture metrics for UMI and non-UMI containing adapters.

4. Application: low-frequency variant calling

One key application of UMIs is to allow for a better signal to noise ratio in calling low-frequency variants. To understand the performance of the Twist UMI adapter system on low-frequency variants, variant calls for the 1% VAF level of a commercially available cfDNA standard were obtained, as well as variant calls for a variety of VAF levels for the Twist cfDNA Pan Cancer Reference Standards. Data was downsampled to 20,000x mean depth over all captured targets, duplex consensus reads were called, and reads were filtered to include only families with at least one molecule from each strand.

In the 1% VAF commercial material, a sensitivity to SNPs of 95-100%was observed. These measurements were based on a relatively small number of sites, and were conducted only at one variant allele frequency, in contrast to the variants from the Twist Pan Cancer Reference standards. Here, similar results were found for the 1% allele fraction, and further show that 50% sensitivity corresponds roughly to an allele frequency of 0.1% for SBS's and 0.25% for indels (based on pileups), based on 20,000x sequencing and the use of duplex consensus reads.

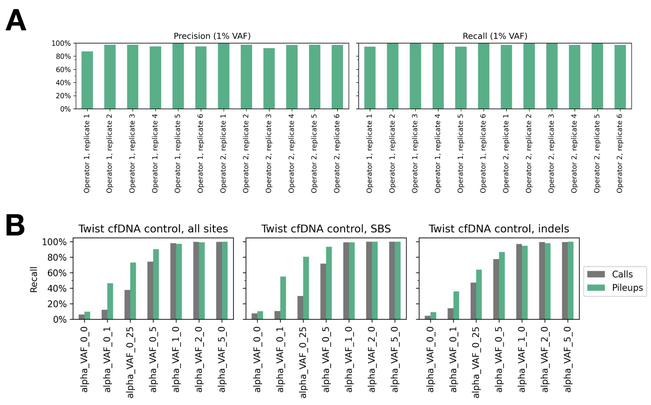


Figure 2: **(A)** Study of precision and recall from a commercially available cfDNA standard at 1% VAF. Two different operators performed 6 replicate captures each, and precision and recall for SBS variants was evaluated. **(B)** Recall over all sites, SBS's and indels alone for duplex consensus reads in the Twist Pan-Cancer cfDNA controls. Pileups represent detection of at least 1 duplex consensus read supporting the variant, while variant calls represent the output from Mutect2.

5. Application: determining library conversion

In library preparation, a key variable that is often difficult to measure is the amount of input DNA that is efficiently converted into sequenceable libraries, or the library conversion efficiency. Ordinarily, it is not possible to get an exact estimate of the total amount of diversity in a sequencing library. However, with the use of UMIs, each individual library molecule can be tagged separately. By counting unique molecules, and comparing the number of unique detected molecules to the number of genome equivalents used to create the library, the efficiency of library conversion can be inferred.

This experiment was performed using sheared NA12878 DNA, patient cfDNA, and the Twist cfDNA Pan Cancer reference standards. To ensure that conversion in a linear range was measured in a linear range, a titration of input masses (10, 20 and 30ng) were investigated. After confirming that the results fit into a linear range, regression equations were used to obtain robust estimates of library conversion efficiency for each sample type.

Twist cfDNA standards consist of both synthetic DNA containing cancer variants, and background DNA derived from a healthy donor. To confirm that synthetic DNA was incorporated into libraries with similar efficiency to background DNA, the number of consensus reads supporting either the reference allele or the expected variant allele at each mass point were identified. The data support equal conversion rates for the synthetic and background DNA, an important condition for ensuring that variants are detected with the desired frequency.

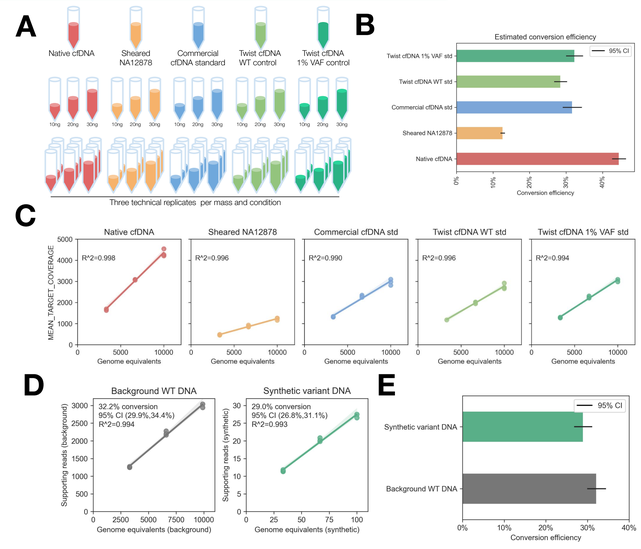


Figure 3: **(A)** Experimental schematic for mass-titration experiment. **(B)** Estimates of library complexity for each sample. **(C)** Correlation of input genomic equivalents to measured non-redundant read counts over target sites. **(D)** Correlation of genome equivalents to observed background and synthetic DNA **(E)** Estimates of conversion efficiency for background vs synthetic DNA in the Twist pan-cancer cfDNA reference standards from the regression analysis shown in D.

6. Application: determining error rates in DNA constructs

In synthetic biology applications, it is important to have a concrete assessment of the total rate of low-frequency errors introduced during DNA handling. As UMIs facilitate extremely low-frequency variant calls, the Twist UMI Adapter System was used to estimate error rates in these applications. Specifically, the rate of errors introduced in human genes that were incorporated either into dumbbell DNA (dbDNA) or standard plasmid constructs were compared. dbDNA constructs have potential advantages in terms of yield, purity, and cost to produce over traditional vectors. Both types of constructs have generally similar error rates and vary based on gene. Error rates were extremely low - between 0.002% and 0.004%, or between 20 and 40 ppm. Overall, the data supports the use of dbDNA constructs as a replacement for plasmids. Furthermore, this establishes that the UMI error rate for bare duplex calls is at most roughly 0.004%, although it is likely considerably lower.

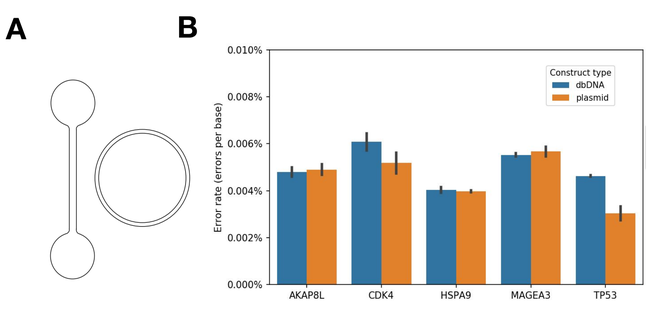


Figure 4: **(A)** Topology of dumbbell (left) and plasmid (right) DNA constructs.
(B) Synthesis and expansion error rates determined using duplex sequencing for dumbbell and DNA constructs. Constructs were sequenced to 80,000x mean depth, and reads were consensus called using bare duplex consensus (i.e. at least one read from each original strand.)

7. Conclusions

Here we describe the design of a new UMI adapter system compatible with Twist's existing next-generation sequencing (NGS) workflow. We describe the design constraints that went into determining the number of UMI sequences necessary for both efficient error correction and avoiding collisions between sequences. In comparisons to non-UMI containing adapters, we show general equivalency for percentage off bait, 80 fold base penalty, and HS library size sequencing metrics.

We demonstrate the utility of UMIs in providing accurate deduplication, using the UMI system to estimate the library conversion efficiency of a number of different input sample types. Additionally, we report allele-specific conversion frequencies for the synthetic background and variant-containing material in the Twist Pan-Cancer cfDNA controls. These results provide a proof of principle for a new assay for library conversion efficiency.

In the synthetic biology space, we show that Twist's UMIs can be used to determine the error rate of different strategies for synthetic DNA constructs. We also use Twist's UMIs in a low-frequency variant calling application to determine the effectiveness of error-correction in distinguishing between true- and false-positive variants. In summary, the Twist UMI Adapter System is effective at collapsing library duplicates and error-correcting reads, and allows for a number of applications that would otherwise be challenging or impossible with standard adapters.

References

Li H, Durbin R, Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, Volume 25, Issue 14, 15 July 2009, Pages 1754–1760,
Schmitt MW, Kennedy SR, Salk JJ, Fox EJ and Loeb LA. Detection of ultra-rare mutations by next-generation sequencing, PNAS, 109(36)14508-14513. 1 August 2012
Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017 Mar;27(3):491-499.
Yu H, Jiang X, Tan KT, Hang L, Patze VI, Efficient production of superior dumbbell-shaped DNA minimal vectors for small hairpin RNA expression, *Nucleic Acids Research*, Volume 43, Issue 18, 15 October 2015, Page e120,

Financial Disclosures: All authors are employees and shareholders of Twist Bioscience